Máster en Ciencias Actuariales y Financieras

TRABAJO FINAL DE MÁSTER

Título: ***Una aplicación de la regresión cuantílica a la estimación de los percentiles de coste total de los siniestros.***

Autoría: Marco Antonio Ospina Ruiz.

Tutoría: Ana María Pérez-Marín, PhD.

Curso académico: 2019 - 2020

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Una Aplicación de la Regresión Cuantílica a la Estimación de los Percentiles del Coste Total de los Siniestros

Trabajo Final de Máster
Máster en Ciencias Actuariales y Financieras
Facultad de Economía y Empresa

Universitat de Barcelona

Autoría: Marco A. Ospina Ruiz
Tutoría : Ana María Pérez-Marín, PhD.

9 de junio de 2020

Resumen

El presente trabajo tratará una aplicación empírica del modelo conocido como *Two Stage Quantile Regression* (Heras *et al* 2018) [5] con el propósito de hacer estimaciones sobre los cuantiles del costo total de siniestros por póliza para una base de datos aseguradora, lo cual proverá información útil acerca de la severidad de las pérdidas monetarias asociadas a las pólizas más riesgosas para la compañía. Dicho modelo se dividirá en dos etapas, la primera se llevará a cabo implementando una regresión logística sobre una variable dicotómica que indicará la presencia de uno o más siniestros en cada póliza, en función de un conjunto de variables explicativas. Dichas variables se incluirán o no en la segunda etapa en base a la significación estadística de los parámetros estimados en la regresión. También se extraeran la probabilidades de que las pólizas reporten al menos un siniestro. En la siguiente etapa se hará uso de esta probabilidad como parámetro para estimar los cuantiles de la distribución de la suma total de cuantías provocadas por los sinietros. Esta metodología será útil para calibrar las estimaciones que deben considerarse en el proceso de suscripción de pólizas en el sector asegurador.

Palabras Clave: Regresión cuantílica, Regresión logística, Modelo de regresión de dos etapas, Riesgo de suscripción, Costo total de siniestros.

Abstract

This paper will address an empirical application of the model known as Two Stage Quantile Regression (Heras et al 2018) [5] with the purpose to make estimates of quantiles of the aggregate claim amount per policy for an insurance database, which will provide useful information about the severity of monetary losses associated with the most risky policies for the company. This model will be divided in two stages, the first one will be by implementing a logistic regression on a dichotomous variable which will indicate the presence of one or more claims in each policy, depending of a set of explanatory variables. These variables may or may not include in the second stage based on the statistical significance of the parameter estimated in the regression. Also, the probabilities of the policies will report at least one claim. In the next stage, this probability will be used as a parameter for estimating the quantiles of the distribution of the total sum of amounts caused by the claims. This methodology will be useful in calibrating the estimates to be considered in the underwriting process of policies in the insurance sector.

Keywords: Quantile regression, Logistic regression, Two stage regression model, Underwriting process, Aggregate claim amount.

El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

.

Índice

1. Introducción

El presente trabajo pretende reproducir la metodología usada en el artículo (Heras *et al* 2018) [5] en el cual se aplica la regresión cuantílica al proceso de *underwriting* en el contexto asegurador, también conocido como riesgo de suscripción. El propósito será estimar los cuantiles del costo total de los siniestros de una base de datos aseguradora, que recoge múltiples pólizas de seguros automovilísticos. Estas pólizas estan segmentadas a través de diversas características que se pueden interpretar como factores de riesgo, que serán de utilidad para explicar el comportamiento de la siniestralidad en la cartera.

El riesgo de suscipción se puede derivar, entre otros factores, de la estimación inexacta de los riesgos asociados a la suscripción de una póliza de seguro, lo que puede ocasionar para la aseguradora que sus costos excedan significativamente las primas recaudadas. Esta falta de precisión en las estimaciones pueden deberse a la inadecuación de las hipótesis de tarificación y a la constitución de provisiones, lo cual está directamente relacionado con el **margen de riesgo**, concepto crucial para los supervisores en el marco de Solvencia II, así como en la *International Financial Report Standard no. 17* (IFRS17).

La gestión del márgen de riesgo por lo tanto es de importancia capital, y es ampliamente discutida por académicos y actuarios activos en la industria ya que no existe una definición o modelo único con el que se pueda llevar a cabo. (Baione y Biancalana 2019) [2]

En nuestro caso se aplicará la metodología *Two Stage Quantile Regression*, o modelo de regresión cuantílica de dos etapas, a una base de datos compuesta también por pólizas automovilísticas con un total de 39075 asegurados, en la cual se pueden observar variables como la edad del conductor, la zona de conducción habitual, el sexo, la antigüedad del permiso de conducir, sus características alimenticias, etc. Nos interesará estimar la probabilidad de que cada una de las pólizas tenga por lo menos un siniestro, en términos de las características anteriormente nombradas.

Los modelos de dos partes en regresión lineal generalizada (GLM) son utilizados en la literatura con el propósito de hacer predicciones sobre la pérdida esperada o el costo total de los siniestros. La metodología central de la que se hará uso en este trabajo considera un **método alternativo**, la aplicación de la regresión cuantílica al proceso de pricing que se ha de implementar en el sector asegurador. Este modelo incorpora

dos etapas, la primera haciendo uso de la regresión logística, también conocida como modelo Logit, sobre una variable aleatoria binaria que determinará si las pólizas han tenido por lo menos un siniestro.

Subsecuentemente se llevará a cabo la regresión cuantílica (QR), que nos dará información acerca de qué tanto riesgo puede representar una póliza individualmente en la cartera, al hacer la estimación de los cuantiles del costo agregado por póliza de los siniestros.

Una de las ventajas de esta regresión es el hecho de que asume muchas menos hipótesis que el método GLM y a su vez es una regresión robusta. [5] Este método no requiere información sobre la distribución de probabilidad de las variables involucradas.

1.1. Objetivos

Se pretende conseguir una mejor estimación de los riesgos asociados al proceso de suscripción de pólizas aseguradoras de las compañías, a través de la estimación de los cuantiles del costo agregado total generado por los siniestros de las pólizas aseguradas. Para conseguirlo se hará uso del modelo de dos etapas que combina tanto la regresión logística como la regresión cuantílica.

La primera etapa tiene como objetivo tanto la **estimación de los parámetros** de las variables explicativas del modelo, así como el **cálculo de las probabilidades** de que las pólizas reporten uno o más siniestros. Para conseguirlo se implementará la regresión cuantílica sobre un conjunto de variables que están recogidas en la base de datos aseguradora. Esta estimación indicará la significación estadística de los parámetros de las variables consideraras.

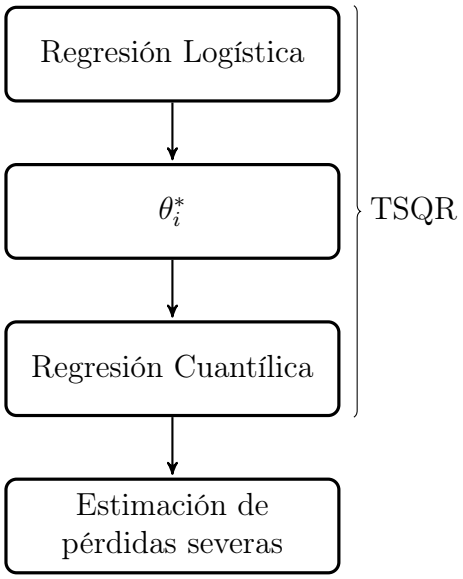
Una vez se haya determinado qué parámetros son significativos se procederá a calcular los niveles de probabilidad a los cuales los parámetros de la regresión cuantílica serán estimados. Dichas pólizas se recojerán en grupos de riesgo que se han de definir más adelante. Esta estimación nos proporcionará información sobre la frecuencia de siniestralidad de la cartera. Estas probabilidades las llamaremos como θ_i^*, donde i es el número de pólizas de la cartera.

Dado que ya se han calculado estas probabilidades se seguirá con la etapa dos del modelo. Allí se implementará la regresión cuantílica para cada una de las probabilidades

consignadas en el vector θ_i^*, habrá un modelo distinto para cada una.

En cada uno de estos modelos obtendremos los parámetros de la regresión implementada y por tanto podemos proceder a **estimar los cuantiles** del costo total de siniestros para cada uno de los grupos de riesgos definidos. De nuevo, el método de regresión nos proporcionará información acerca de las variables cuyos parámetros han sido significativos estadísticamente.

Esta información será util para hacer estimaciones sobre posibles pérdidas severas monetarias, en el sentido en el que la regresión cuantílica nos puede dar información acerca de qué pólizas pueden representar más riesgos para la compañía. Todo este proceso se resume en el siguiente diagrama:



Subsecuentemente se procederá a estimar el costo total de siniestros para cada una de las tarifas que establecerá cada uno de los grupos de riesgos que se implementará en el modelo, combinando las estimaciones de la regresión logística y la regresión cuantílica.

Como se ha descrito en la sección anterior, este tipo de modelos pretende proveer una metodología razonable en la búsqueda de la estimación de posibles pérdidas

monetarias severas asociadas al ejercicio del aseguramiento, con el fin de hacer una buena gestión de la tarificación y las provisiones técnicas, que tendrán un impacto significativo y directo sobre el margen de riesgo que establecen los reguladores.

1.2. Revisión bibliográfica

El riesgo de suscripción juega un papel trascendental en el marco de la directiva de *Solvencia II* a nivel europeo, en el contexto del cálculo del capital de solvencia obligatorio para el riesgo de suscripción distinto del de vida. Por tanto resulta relevante hacer una adecuación precisa de las hipótesis de tarificación y constitución de provisiones, en este sentido serán **relevantes** las técnicas con las que se pretende hacer estimaciones sobre el costo total de los siniestros en las pólizas de seguros.

El modelo que se implementará en este trabajo es el modelo de regresión cuantílica en dos etapas, el cual fue introducido en el artículo Heras *et al* 2018, en el cual se busca estimar los cuantiles del costo total de los siniestros, los cuales pueden proporcionar información crucial para evaluar el riesgo asociado a cada póliza, como se ha remarcado antes.

Dicha metodología está basada en el método de regresión cuantílica, el cual fue presentado en el artículo *Regression Quantiles* de los autores Koenker y Bassett (1978) [6] El método se estableció con el objetivo de hacer la estimación de funciones cuantílicas condicionales, es decir, modelos en los cuales los cuantiles de la distribución condicional de la variable de respuesta vienen dados en función de variables explicativas. [6]

Varios autores han hecho uso de estos métodos en el campo del *pricing* en el sector asegurador, de los cuales cabe destacar: Baione y Biancalana (2019) [2] los cuales implementan el método de la regresión cuantílica y los modelos lineales generalizados para el cálculo de primas basadas en cuantiles. Desarrollaron un proceso de cálculo de primas basado en un modelo de dos partes usando la regresión cuantílica y un modelo Gamma GLM respectivamente, con el propósito de estimar la severidad de los cuantiles del coste de los siniestros.

Kudryavtsev (2009) [1], Se planeta la regresión cuantílica como un método alternativo a los modelos de regresión lineales generalizados en el proceso de estimación de las primas en el contexto asegurador. Evidencia las ventajas que tiene este método sobre los métodos tradicionales. Señalan que los métodos de regresión tradicional tienen algunas desventajas, particularmente su sensibilidad a las hipótesis que se hacen acerca de la distribución de las variables aleatorias involucradas en el proceso, lo cual limita su aplicabilidad.

También muestra que la calidad de los parámetros de la regresión cuantílica es aproximadamente la misma, o en algunos casos mucho mejor, que los obtenidos por los métodos tradicionales.

Pitselis (2013) [7] Desarolló una relación entre la teoría de la credibilidad en los seguros y los cuantiles. Mostó cómo los cuantiles pueden ser introducidos en el modelo clásico de credibilidad de Bühlmann (1967) y en el modelo de regresión de credibilidad de Hachemeister (1975). Basó su estimación de la credibilidad en lo que se denomina la función de influencia para el p-cuantil.

Portnoy [8] Aquí se aplica la regresión cuantílica presentada por Koenker y Bassett (1978) en el uso de la desviación absoluta en vez de la desviación al cuadrado para medir el error de ajuste de los datos. El método Schette de graduación es un caso especial de este tipo de métodos, y actualmente no se usa mucho en el contexto actuarial. Lo que se pretende en el artículo es demostrar la efectividad de ese método haciendo estimaciones de las curvas dadas por los cuantiles condicionales.

2. Metodología

2.1. Modelo de Regresión Lineal Generalizado

Los modelos de regresión pretenden explicar el comportamiento de una variable que se llamará **variable de *respuesta***, que se puede entender como la variable dependiente de interés en el modelo, en términos de un conjunto de **variables *predictoras***, que juegan el papel de las variables independientes del mismo.

En los modelos de regresión lineal generalizados (GLM) se puede expresar la esperanza matemática de la variable dependiente en función de una combinación lineal de variables independientes. (Frees 2010) [4]

Se suele usar la notación $E[y_i] = \mu_i$, con $i \in \mathbb{N}$, para la esperanza de la variable independiente y_i, y $\eta_i = x_i'\beta$ lo que se denomina *componente sistemático* en términos de sus variables predictoras. Aquí x_i' es un vector transpuesto que contiene cada una de estas variables y β cada uno de los parámetros de la estimación.
Además, existe una expresión que relaciona la esperanza matemática con el componente sistemático de la siguiente manera:

$$\eta_i = x_i'\beta = g(\mu_i)$$

La función g que depende de μ_i se conoce como la función de enlace, o función *link*. Con respecto a la distribución de la variable de respuesta, que servirá para especificar el modelo GLM, se hará uso de una **distribución de la familia exponencial**.

Dada una variable de respuesta y, se puede especificar un modelo general de la siguiente manera:

$$f(y) = c(y, \phi)exp\left\{\frac{y(\theta) - a(\theta)}{\phi}\right\},\ g(\mu) = x'\beta$$

De aquí tenemos que la escogencia de la función $a(\theta)$ determinará la **distribución** de la variable de respuesta, mientras que $g(\mu)$ será la respectiva función link y establecerá la relación que tiene la media con las variables predictoras x, donde g es una función diferenciable y monótona.

Consideremos la variable de respuesta y que toma valores $y = 0$ o $y = 1$, que puede indicar si la póliza en cuestión ha tenido o no siniestros, tendremos que $y \sim B(1, \pi)$, donde π es la probabilidad de que $y = 1$, y por tanto $E(y) = \pi$ y $Var(y) = \pi(1 - \pi)$.

La distribución de Bernoulli es un caso particular de la distribución Binomial, con $n = 1$ (de Jong y Heller 2008) [3] la cual es un miembro de la familia exponencial, y será nuestro componente sistemático de interés, y su GML está dado por:

$$y \sim B(1, \pi)$$
$$g(\pi) = x'\beta$$

los cocientes $\pi/(1 - \pi)$ conocidos como ***odds ratios*** indican proporcionalmente qué tan probable es la ocurrencia de un siniestro comparado con su no ocurrencia. Esta expresión corresponde a la **regresión logística**.

Se especifica la función de enlace (*link*) de la regresión logística como el logarítmo de los odds ratios, en términos de las variables independientes de la siguiente manera:

$$g(\mu) = ln\frac{\pi}{1-\pi} = x'\beta$$

$$\pi = \frac{e^{x'\beta}}{1+e^{x'\beta}}$$

Por tanto tenemos que los GML consisten en tres elementos:

- Se puede observar que la distribución de la variable de respuesta $f(y)$ pertencece a una familia exponencial de distribuciones de probabilidad.
- La función (o transformación) que se aplica sobre la media $g(\mu)$ está relacionada linealmente a las variables predictoras x en el vector x'.
- Una **función de media** como la función inversa de la función de enlace: $\mu = g^{-1}(x'\beta)$.

2.2. Correción por exposición

Con el propósito de implementar la regresión logística para estimar la probabilidad de ocurrencia de los siniestros, hemos de tener en cuenta que las pólizas por lo general han tenido una cierta exposición temporal al riesgo t, como lo indica de Jong y Heller (2008) [3], podremos tener en cuenta este efecto sobre dichas probabilidades de la siguiente manera:

Si consideramos $0 < t \leq 1$, tendremos que la probabilidad de ocurrencia de al menos un siniestro se reducirá proporcionalmente a causa de dicha exposición, por lo que nos veremos en la necesidad de modificar la función de enlace correspondiente a la regresión logística para captar adecuadamente este efecto.

Si π es la probabilidad de ocurrencia de siniestros a lo largo de un periodo, definamos $\pi^* = t\pi$, por tanto:

$$ln\frac{\pi^*/t}{1-\pi^*/t} = x'\beta$$
$$t\frac{e^{x'\beta}}{1+e^{x'\beta}} = \pi^*$$

Dicha correción se implementará posteriormente en SAS en el momento de llevar a cabo la estimación de la regresión logística

2.3. Regresión Cuantílica

Una vez especificado el modelo de regresión lineal generalizado, se puede hablar del modelo de regresión cuantílica (QR), el cual tiene diversas aplicaciones en el proceso de *pricing* en el contexto asegurador así como en el proceso de underwriting [5].

A modo de introducir el concepto de regresión cuantílica consideremos lo siguiente: Normalmente estamos interesados en un modelo de regresión que estime la esperanza matemática condicional de una variable aleatoria S para un cierto cuantil fijo m en términos de una expresión de la siguiente manera:

$$E[S|x] = x\beta_m$$

Este tipo de estimación es el que suele considerarse haciendo uso de los GLM para el cálculo de la prima pura, como la esperanza matemática de la pérdida esperada en términos de ciertos factores de riesgo. Ahora, así mismo, podríamos estar interesados en la mediana de la distribución:

$$Mediana[S|x] = x\beta_{0,5}$$

Y en general, se puede considerar los cuantiles τ, tales que $\tau \in (0;1)$:

$$Cuantil_\tau[S|x] = x\beta_\tau$$

Observamos que de la misma manera los cuantiles condicionales pueden estimarse en términos de un conjunto de factores de riesgo, lo que permite, para un cierto nivel de probabilidad dado, estimar la cuantía máxima total de los siniestros para cada póliza perteneciente a cierto grupo de riesgo. (Heras *et al* 2018) [5].

Inversa generalizada: Por lo tanto, para definir el concepto de función cuantil, se debe primero definir la *inversa generalizada de una distribución de probabilidad*. Dada una función creciente $T : \mathbb{R} \mapsto \mathbb{R}$, la inversa generalizada se define como:

$$T^{-1}(y) = inf\{x \in \mathbb{R} : T(x) \geq y\}$$

Función cuantil: Dada una función de distribución F, la inversa generalizada de F se llama función cuantil de F, para $\tau \in (0, 1)$. El cuantil τ de F viene dado por:

$$Q_s(F) = F^{-1}(\tau) = inf\{x \in \mathbb{R} : F(x) \geq \tau\}$$

En general, el método de la regresión cuantílica tiene algunas propiedades preferibles sobre los GLM [5]:

- No asume hipótesis sobre la distribución de la variable de respuesta.
- El método de regresión cuantílica es robusto frente a valores atípicos.

Lo que se pretende es establecer la relación lineal entre los cuantiles de la variable de respuesta S_i y sus variables predictorias x, esta relación viene dada por la siguiente expresión.

$$Q_{S_i}(\theta \mid x_i) = x_i\beta_\theta$$

Obteniendo los parámetros β_θ a partir de la optimización lineal del siguiente problema de minimización:

$$\min_{\beta_\theta} \frac{1}{n} \left\{ \theta \sum_{i:s_i \geq x\beta_\theta} |s_i - x_i\beta_\theta| + (1-\theta) \sum_{i:s_i < x\beta_\theta} |s_i - x_i\beta_\theta| \right\}$$

Donde $(s_i, x_i), i = 1, ..., n$, además, $Q_{S_i}(\theta \mid x_i)$ se puede interpretar como el cuantil condicional de la variable aleatoria S_i para cierta probabilidad θ dado un vector de variables predictoras x_i. Se ha hecho uso de la notación extraida de A. Heras *et al* 2018 [5].

2.4. Modelo en dos etapas

Se han especificado las características de los modelos de regresión logística y regresión cuantílica, con el propósito de implementar el modelo de dos etapas, para estimar la distribución de probabilidad del costo total de los siniestros asegurados. Este método es valioso en el sentido en el que permite, para siniestros de una cuantía atípicamente elevada, identificar los factores de riesgos con los cuales está relacionada y por tanto, qué asegurados son más propensos a tener una mayor siniestralidad en términos de su costo (A. Heras *et al* 2018) [5].

Dado S_i que denota el costo total de los siniestros para un asegurado i, con $i = 1, \ldots, I$ donde I denota el número total de asegurados que componen la cartera aseguradora, y sea N_i el número de siniestros de cada asegurado i, lo que se pretende es estimar $Q_{S_i}(\theta \mid x_i)$, teniendo en cuenta que $\theta = F_{S_i}(s_i|x_i)$ es la probabilidad acumulada hasta el monto de la póliza s_i, condicionada a las variables explicativas x_i.

Se tiene entonces la siguiente descomposición (A. Heras *et al* 2018) [5], dado que:

$$P(S_i \leq s_i|x_i) = P(N_i = 0|x_i) + P(N_i > 0, S_i \leq s_i|x_i)$$

y también

$$P(N_i > 0, S_i \leq s_i|x_i) = P(N_i > 0|x_i)P(S_i \leq s_i|N_i > 0, x_i)$$

Se tiene que

$$P(S_i \leq s_i|x_i) = P(N_i = 0|x_i) + P(N_i > 0|x_i)P(S_i \leq s_i|N_i > 0, x_i)$$

Esta expresión es equivalente a la siguiente, en términos de las funciones de distribución acumuladas de cada una de las variables aleatorias correspondientes:

$$F_{S_i}(s_i|x_i) = F_{N_i}(0|x_i) + (1 - F_{N_i}(0|x_i))F_{S_i|N_i>0}(s_i|x_i) \tag{1}$$

Ahora, se define la probabilidad de que la póliza i no tenga siniestros, y se denotará por $p_i = F_{N_i}(0|x_i)$, por tanto se establece:

$$\theta_i^* = \frac{\theta - p_i}{1 - p_i}$$

De la relación obtenida anteriormente (1), se puede inferir que $F_{S_i}(s_i|x_i) = \theta$ es equivalente a $F_{S_i|N_i>0}(s_i|x_i) = \theta_i^*$ (A. Heras *et al* 2018) [5]. De lo anterior se tiene

que el cuantil θ de S_i equivale al cuantil θ_i^* de $S_i|N_i > 0$, es decir,

$$Q_{S_i}(\theta|x_i) = Q_{S_i|N_i>0}(\theta_i^*|x_i)$$

Por lo tanto, en la primera etapa del modelo se estimará la probabilidad de que la póliza i no reporte ningún siniestro, la cual se había denotado p_i. Para conseguirlo se usa el modelo de regresión logística (o modelo logit):

$$\begin{aligned} logit(1 - F_{N_i}(0|x_i)) &= ln\left(\frac{1-p_i}{p_i}\right) \\ &= x_i'\beta \end{aligned}$$

Consecuentemente en el etapa 2 se aplica la regresión cuantílica para cada póliza en su respectivo grupo de riesgo, de acuerdo a la probabiliad estimada de presentar al menos un siniestro.

Se ha escogido un nivel de probabilidad $\theta = 0{,}95$ porque es el nivel al que se suele considerar en la literatura (De Jong y Heller (2008) [3], Kudryavtsev (2009) [1], Frees (2010) [4]). Para obtener una tarifa multiplicativa la variable de respuesta de la regresión cuantílica es el logarítmo del costo agregado de siniestros por póliza $Log(S_i)$ condicionado al evento de tener uno o más siniestros.

En cuanto se implementa la regresión cuantílica para esta transformación sobre la variable de respuesta, se obtienen los cuantiles de variable de respuesta original S_i en términos de [5]:

$$Q_{S_i|N_i>0}(\theta_i^*|x_i) = exp(x_i'\gamma_{i,\theta_i^*})$$

Por lo tanto, este modelo de regresión logística permite estimar los cuantiles condicionales del coste total de siniestros dadas unas variables explicativas. Dado un nivel de probabilidad, se puede estimar el importe total máximo de siniestros de una póliza dada en alguno de los grupos de riesgo. Por consiguiente, la regresión cuantílica puede ser un método apropiado en términos de la segmentación de riesgo. (Heras *et al* 2018) [5]

En la segunda etapa esto se llevará a cabo con el paquete *quantreg* para R, implementado por R. Koenker [1]

[1] https://cran.r-project.org/web/packages/quantreg/index.html

3. Aplicación Empírica

3.1. Descripción de la base de datos aseguradora

Para la aplicación empírica se hará uso de una base de datos aseguradora [2] que recoge un total de 39075 pólizas automovilísticas. Está compuesta de datos que caracterizan el perfil de riesgo de cada uno de los asegurados, como lo son: el costo total de siniestros por póliza y el costo de cada uno de los siniestros, la edad, el sexo, la marca de fabricación del vehículo, la antigüedad del permiso de conducción, la exposición temporal de cada póliza, el año de fabricación del vehículo, entre otros.

Por otro lado, la siniestralidad de la base de datos se distribuye de la siguiente manera:

Frecuencia de los siniestros						
Siniestros	0	1	2	3	4	5
N. de pólizas	33386	4957	646	78	7	1
N. total de siniestros	0	4957	1292	234	28	5
% De pólizas que reportan siniestros	85,44 %	12,68 %	1,65 %	0,20 %	0,018 %	0,0025 %

Cuadro 1: *Frecuencia de los siniestros de la muestra. Se reportaron un total de 6516 sinietros.*

Como es común para las carteras de siniestros automovilísticos, estamos ante lo que se conoce como *zero-inflated model*, en el sentido en el que el dato más frecuente es la ausencia de siniestros. Se tiene un total del 85,44 % de pólizas que no han reportado sinietros.

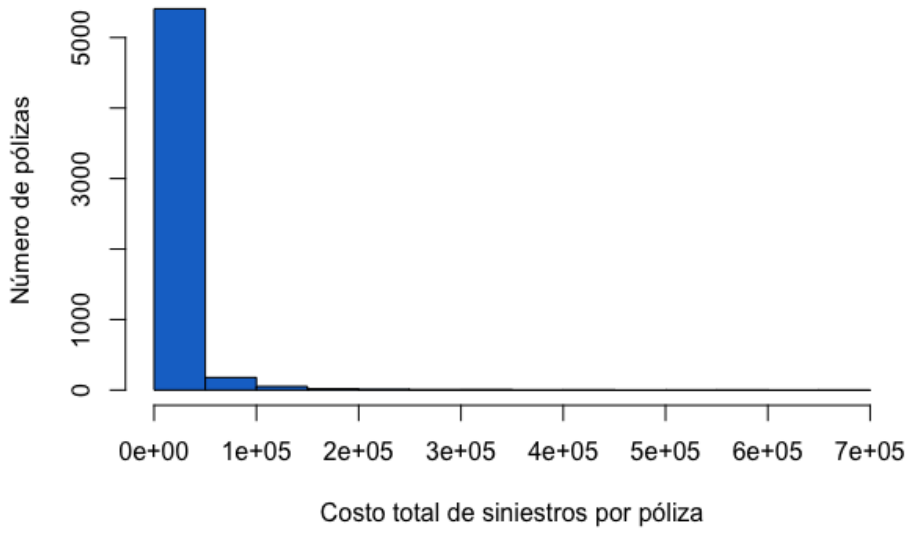
[2] Fuente: A. Charpentier, UQAM (Université du Québec à Montréal)
https://www.r-bloggers.com/r-for-actuarial-science/

Ha de resaltarse que para cada póliza se define el costo total de los siniestros como la suma de la cuantía de cada uno de los siniestros reportados por cada asegurado. El histograma (*figura 1*) representa la distribución de las pólizas que han reportado por lo menos un siniestro, es decir, 6516 pólizas de las 39075 totales.

Estadísticos básicos	Min.	Media	Max.	Desviación Típica
Coste total de siniestros.	0	1326	665752	12934.05
Edad del asegurado.	18	45,59	87	12,05
Antiguedad permiso de conducción.	0	22.15	77	9.11
Año matriculación vehículo.	1945	1995	2003	4.44
Número de siniestros por póliza.	0	0,167	5	$0,43$

Cuadro 2: *Estadísticos básicos para diferentes variables que componen la base de datos aseguradora.*

Figura 1: Histograma del costo total con al menos un siniestro.



3.2. Variables involucradas en el modelo

La base de datos aseguradora recoge información relevante acerca de cada uno de los asegurados que forman parte de su cartera. Algunos de estos datos pueden interpretarse como factores de riesgo al momento de reportar siniestralidad en cada una de las pólizas. Estos factores de riesgo serán las **variables explicativas** que se tendrán en cuenta en el modelo que se va a implementar.

Por otro lado tendremos también información acerca de el importe total de los siniestros y el número de siniestros por póliza. Con esta información vamos a definir las **variables de respuesta** del modelo. Lo anterior se definirá a continuación.

Variables de respuesta del modelo:

- *cout*: Costo total de siniestros por póliza.
- *claim*: Será la variable que indicará si, para cada póliza, se reportó o no por lo menos un sinietro (Sí = 1, No=0). Se ha definido a través de la variable *nbsin* que reporta el número total de siniestros por póliza en la base de datos.

Variables explicativas del modelo:

- *age*: Edad del asegurado en años.
- *sex_g*: Sexo del asegurado (Mujer = 0, Hombre = 1). En esta base de datos hay 15846 mujeres y 23229 hombres.
- *duree_permis*: Antigüedad del permiso de conducción del asegurado.
- *anee_vehicule*: Año de matriculación del vehículo.
- *marque_voiture*: Marca de fabricación del vehículo.

Para definir los grupos de riesgo que se van a implementar en el modelo, se agruparon algunas de las variables anteriores de la siguiente manera:

Variables agrupadas:

- *age_g*: Se han definido 4 grupos: jóvenes adultos (18-34 años), adultos (35-51 años), adultos mayores (52-68 años) y vejez (69-87 años). Cada grupo tiene una separación uniforme de 17 años.
- *duree_permis_g*: Se han definido 5 grupos: grupo 1 (0-14 años), grupo 2 (15- 29 años), grupo 3 (30-44 años), grupo 4 (45-60), grupo 5 (61 años o más). Cada grupo tiene una separación uniforme de 15 años.
- *anee_vehicule_g*: Se han definido 6 grupos ascendentemente desde el más antiguo al más reciente.

Al implementarse el modelo de regresión logística, del cual veremos los resultados en la siguiente sección, se han descartado algunas variables que no se han de implementar en la etapa de la regresión cuantílica. Las variables *duree_permis_g*, *anee_vehicule_g* y *marque_voiture* se han descartado por su falta de significación estadística, sus p-valores correspondientes se pueden apreciar en el Cuadro 11 en la sección de anexos.

Una vez se han descartado las variables anteriores, se reestimó el modelo de regresión logística con las variables que sí resultaron tener un efecto significativo. El Cuadro 3 resume las variables que se usaron en la primeta etapa. En el Cuadro 4 se resume el modelo con sus respectivos parámetros estimados y p-valores.

Variable	Descripción
sex_g	Sexo del asegurado (1=Hombre, 0=Mujer)
age_g	Edad del asegurado: 1 (más jóven), 2 ,3 ,4)
claim	Ocurrencia del siniestro (1=sí, 0=no)
exposition	Número de años de la póliza (entre 0 y 1)

Cuadro 3: *Resumen de las variables con las que se ha estimado el modelo de regresión logística final.*

3.3. Resultados del modelo en dos etapas

Para la implementación del modelo se han creado 2 x 4 = 8 grupos de riesgo disintos, en cuanto al sexo del asegurado y a su grupo de edad correspondiente. La variable de respuesta para **la regresión logística** será la variable dicotómica *claim*, y el modelo de regresión logística estará especificado de esta manera: (Heras *et al* 2018) [5]

$$ln((1-p_i)/p_i) = \beta_0 + \beta_1(sex\_g_i) + \beta_2(age\_g_i)$$

Aquí, $(1-p_i)$ es la probabilidad de que la póliza i haya reportado al menos un siniestro. Los resultados obtenidos de la implementación en SAS [3] se pueden observar en el Cuadro 4.

	Estimación	Std. Error	z value	$Pr(>\|z)$	
(Intercept)	-1.7913	0.0800		<.0001	***
sex_g: 0	-0.1827	0.0310		<.0001	***
sex_g: 1	0.0000	0.0000		-	
age_g: 1	0.6312	0.0873		<.0001	***
age_g: 2	0.5104	0.0829		<.0001	***
age_g: 3	0.2600	0.0855		0.0024	***
age_g: 4	0.0000	0.0000		-	

Cuadro 4: *Resultados de la estimación utilizando el modelo de regresión logística.*

[3] El procedimiento implementado ha sido PROC GENMOD, Nelder y Wedderburn (1972) https://support.sas.com/rnd/app/stat/procedures/genmod.html

Por otro lado, aquí cabe destacar que la estimación se ha hecho teniendo en cuenta la correción por exposición que se ha descrito en la sección 2.2. a través de la modificación de la función de enlace allí descrita. La correción se ha implementado en *SAS* teniendo en cuenta la variable *exposition*.

Del Cuadro 4 podemos **interpretar los coeficientes** estimados por la regresión. el coeficiente -0.1827 (*sex_ g:0*) nos indica que es más probable que los hombres sean más propensos a reportar siniestralidad en la cartera.

En el caso de las categorías de edad, los grupos de riesgo que agrupan a los asegurados más jóvenes (1 y 2) serán más propensos a reportar siniestralidad en la cartera que los grupos de edad más avanzada (3 y 4).

También se ha estimado [4] a partir del modelo anterior, la probabilidad de que las pólizas reporten cero siniestros p_i (Cuadro 5), para cada una de las 8 tarifas que corresponden a los grupos formados descritos anteriormente.

El cálculo de θ_i^*, el cual nos dará los diferentes niveles de probabilidad a los cuales los parámetros de la regresión cuantílica serán estimados para cada grupo de riesgo se deriva de la fórmula (2). Dicho cálculo se lleva a cabo en términos de p_i, la probabilidad de que se hayan reportado cero siniestros, este valor se ha estimado en la regresión cuantílica. La expresión que los relaciona está dada por [5]:

$$\theta_i^* = \frac{\theta - p_i}{1 - p_i} \tag{2}$$

La regresión cuantílica se estimará para cada uno de los tipos de pólizas agrupados indicados anteriormente, de acuerdo a su probabilidad p_i, y aquí la variable de respuesta será *cout*, la suma del coste total de los siniestros por póliza, condicionada a que por lo menos haya un siniestro. (A. Heras *et al* 2018) [5]

[4] La estimación se ha hecho a través del procedimiento PROC PLM
`https://support.sas.com/rnd/app/stat/procedures/plm.html`

Grupo	N. de pólizas	N. de siniestros	p_i	θ_i^*
1 (1 & 1)	4058	747	0.76135	0.79048
2 (1 & 2)	10983	2133	0.78261	0.77000
3 (1 & 3)	6848	1006	0.82220	0.71879
4 (1 & 4)	1340	173	0.85709	0.65013
5 (0 & 1)	3481	558	0.79295	0.75852
6 (0 & 2)	8688	1358	0.81208	0.73393
7 (0 & 3)	3373	510	0.84735	0.67245
8 (0 & 4)	304	31	0.87804	0.59004

Cuadro 5: *Probabilidades estimadas a través de la regresión logística, la columna p_i representa la probabilidad de que no se haya reportado ningún siniestro, y la columna θ_i^* la probabilidad que servirá como cuantil modificado para la regresión cuantílica.*

Para la estimación se implementará el siguiente modelo en R:

$$Q_{S_i|N_i>0}(\theta_i^*|x_i) = exp(\gamma_{\theta_i^*,0} + \gamma_{\theta_i^*,1}(sex\_g_i) + \gamma_{\theta_i^*,2}(age\_g_i)) \tag{3}$$

En esta ecuación S_i representa a la variable de respuesta, el costo agregado de siniestros para cada póliza i, y θ_i^* será la probabilidad estimada en el paso anterior. conservando las mismas variables explicativas que habíamos definido para el modelo de regresión logística. El resultado de la implementación está consignado en el Cuadro 6.

Se puede notar que a pesar de que los coeficientes de la regresión logística para la primera etapa del método fueron en general muy significativos, esto mismo no sucede para la segunda etapa, la regresión cuantílica, esto se discutirá en la próxima sección. Los resultados completos de la regresión se pueden apreciar en la sección 5 de anexos.

Modelo	1	2	3	4	5	6	7	8
(Intercept)	8.16124 ***	7.96013 ***	7.56249 ***	7.22255 ***	7.90644 ***	7.66109 ***	7.31128 ***	6.94291 ***
sex_g								
1	-0.02032	-0.03658	-0.02302	-0.00849	-0.06321	-0.02410	0.01596	-0.03435
age_g								
1	-0.33944 **	-0.28457 **	-0.15200	-0.09199	-0.26538 **	-0.15819	-0.11445	-0.00707
2	-0.15093	-0.11385	-0.05422	-0.03424	-0.11352	-0.07988	-0.04539	0.01950
3	-0.30463	-0.32183	-0.16310	0.02405	-0.27301	-0.22643	-0.04746	0.12724

Cuadro 6: *Coeficientes de la regresión cuantílica. Cod. Signif. : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1*

Grupo	N. de siniestros	N. de pólizas	p_i	LG + QR Cuantía siniestro a $\theta = 0{,}95$
1 (1 & 1)	654	4058	0.76135	2444,20
2 (1 & 2)	1810	10983	0.78261	2199,22
3 (1 & 3)	889	6848	0.82220	1880,84
4 (1 & 4)	155	1340	0.85709	1495,60
5 (0 & 1)	496	3481	0.79295	2081,95
6 (0 & 2)	1198	8688	0.81208	1810,44
7 (0 & 3)	458	3373	0.84735	1497,09
8 (0 & 4)	29	304	0.87804	1723,09

Cuadro 7: *Costo total estimado de siniestros para cada grupo de riesgo a un nivel de* 0,95*. La columna 5 contiene los resultados de la combinación del modelo de la regresión logística (LG) y la regresión cuantílica (QR). Su cálculo se obtiene a partir de la ecuación (3).*

La interpretación de los coeficientes para la regresión cuantílica (Cuadro 6) es muy similar a la que se hace en la regresión de los mínimos cuadrados ordinarios. Aquí en vez de predecir la media de la variable dependiente, que en nuestro caso es *cout*, la suma del coste total de siniestros, la regresión cuantílica busca los **cuantiles** de la misma.

El Cuadro 7 resume la estimación de los cuantiles para cada grupo de riesgo (LG + QR Cuantía siniestro a un nivel de $\theta = 0{,}95$). Estos cuantiles han sido calculados agrupando las cuantías de los siniestros correspondientes a cada grupo. Una vez agrupadas se calculó el cuantil al nivel de probabilidad θ_i^* que fue estimado anteriormente (Cuadro 5). Los valores obtenidos se discutirán en la sección 5.

3.4. Resultado de la estimación a niveles de percentil θ distintos

Se analizó el impacto que tiene sobre el modelo considerar niveles de percentil distintos a $\theta = 0,95$ y $\theta = 0,975$. Se reestimó el modelo de regresión logística (etapa 1) para ambos niveles. Los resultados obtenidos sobre θ_i^* se pueden observar en el Cuadro 8. Este será el nuevo nivel de probabilidad que servirá como cuantil modificado para la regresión cuantílica (etapa 2).

Grupo	p_i	$\theta^*_{0,9}$	$\theta^*_{0,975}$
1 (1 & 1)	0.76135	0.58097	0.89524
2 (1 & 2)	0.78261	0.53999	0.88499
3 (1 & 3)	0.82220	0.43758	0.85939
4 (1 & 4)	0.85709	0.30026	0.82506
5 (0 & 1)	0.79295	0.51703	0.87926
6 (0 & 2)	0.81208	0.46785	0.86696
7 (0 & 3)	0.84735	0.34490	0.83623
8 (0 & 4)	0.87804	0.18007	0.79502

Cuadro 8: *Reestimación de las probabilidades a partir de los niveles* $\theta = 0{,}9$ *y* $\theta = 0{,}975$

Al implementarse la regresión cuantílica se han obtenido los siguientes resultados:

Modelo	1	2	3	4	5	6	7	8
(Intercept)	7.00364 ***	6.95443 ***	6.65574 ***	6.26232 ***	6.88563 ***	6.69547 ***	6.31962 ***	5.48510 ***
sex_g								
1	-0.02310	-0.06290	-0.06988	-0.15140	-0.07452	-0.06448	-0.12015	-0.01018
age_g								
1	-0.10470	-0.19332	-0.25485	-0.38840	-0.19004	-0.21709	-0.24641	-0.28853
2	-0.07197	-0.12434	-0.19831	-0.32684	-0.12758	-0.11562	-0.21400	-0.23433
3	-0.08673	-0.16112	-0.20704	-0.41198	-0.15472	-0.15467	-0.24061	-0.25862

Cuadro 9: *Coeficientes de la regresión cuantílica reestimada para* $\theta = 0{,}9$

Modelo	1	2	3	4	5	6	7	8
(Intercept)	10.11885 ***	10.09209 ***	8.94579 ***	8.61423 ***	9.84269 ***	9.39812 ***	8.69076 ***	8.17855 ***
sex_g								
1	-0.02266	-0.11369	-0.09954	-0.11698	-0.04092	-0.10368	-0.11302	-0.01561
age_g								
1	-0.44085	-0.74560	-0.20786	-0.31006	-0.64340	-0.41550	-0.25211	-0.33735
2	-0.24800	-0.44101	0.19790	-0.10988	-0.36951	-0.15487	0.01409	-0.23433
3	0.02923	-0.00075	0.48888	0.17347	0.09456	0.25945	0.28906	-0.25862

Cuadro 10: *Coeficientes de la regresión cuantílica reestimada para* $\theta = 0{,}975$

En este caso se puede observar (Cuadro 9 y 10) que ninguno de los parámetros de la regresión cuantílica ha resultado ser significativo. Se evidencia que el modelo es muy sensible a la escogencia del nivel θ para la significación estadística de los parámetros de la estimación de la regresión cuantílica (segunda etapa).

4. Aplicación de la metodología de regresión cuantílica en dos etapas para el cálculo de la prima pura

En el artículo (A. Heras *et al* 2018) [5] se ha propuesto una aplicación sobre el modelo de regresión cuantílica en dos etapas para el cálculo de la prima pura, este método es llamado *Principio de la Prima Cuantílica* .

Lo que se propone es usar la diferencia sobre un cuantil θ prestablecido del costo total de siniestros, $Q_{S_i}(\theta)$ y su valor esperado $E(S_i)$. La prima resultante podrá calcularse entonces como una combinación convexa de $Q_{S_i}(\theta)$ y $E(S_i)$: (Heras *et al* 2018) [5]

$$P_i = \alpha Q_{S_i}(\theta) + (1-\alpha)E(S_i)$$

α se conoce en la literatura como el *factor de recargo* asociado al cálculo de la prima pura. Este principio satisface propiedades como la del no exceso, la consistencia o invarianza a las translaciones y un recargo de seguridad no negativo.

Por consiguiente, el Principio de la Prima Cuantílica tiene ventajas sobre los principios tradicionales con los que se suele calcular la prima pura, como lo son: El principio del Valor Esperado,

$$P_i = (1+\alpha)E(S_i)$$

El cual no satisface las propiedades anteriormente dichas. También el Principio de la Varianza:

$$P_i = E(S_i) + \alpha Var(S_i)$$

Y el Principio de la Desviación Estándar,

$$P_i = E(S_i) + \alpha\sqrt{Var(S_i)}$$

Estos dos principios tienen propiedades estadísticas más adecuadas que el Principio del Valor Esperado, pero tampoco satisfacen las propiedades de no exceso, la consistencia o invarianza a las translaciones y un recargo de seguridad no negativo.

Por tanto, calcular la prima pura como se ha especificado en (A. Heras *et al* 2018) [5], se hace uso de la siguiente formulación:

Asumiendo que $N_i \geq 0$,

$$\begin{aligned} E(S_i) = E(E(S_i|N_i)) &= Prob(N_i = 0)E(S_i|N_i = 0) + Prob(N_i > 0)E(S_i|N_i > 0) \\ &= Prob(N_i > 0)E(S_i|N_i > 0) \end{aligned}$$

En la **primera etapa** se estima la probabilidad $Prob(N_i > 0) = 1 - p_i$ haciendo uso de la regresión logística. En la **segunda etapa** se estima $E(S_i|N_i > 0)$ usando un modelo de regresión gamma.

5. Discusión

Como se pretendía inicialmente, se logró reproducir la metodología del artículo (A. Heras *et al* 2018) [5]. Los resultados obtenidos mediante la aplicación empírica del modelo en dos etapas arrojaron una estimación significativa de la gran mayoría de los coeficientes β de las variables explicativas en la regresión logística. Con ello se procedió a calcular la probabilidad de que cada uno de los 8 grupos de riesgo establecidos no reportasen ningún siniestro.

Por otra parte, se ha verificado que la regresión cuantílica, dotada de propiedades estadísticas convenientes en este tipo de implementaciones, como lo son su falta de necesidad de una hipótesis sobre la distribución del costo total de siniestros y su robustez frente a las colas, puede resultar ser una herramienta muy útil para tener una mejor aproximación sobre la severidad de las pérdidas asociadas a cada una de las pólizas de la cartera, ya que nos permite identificar, póliza a póliza, aquellas que pueden representar más riesgo.

Sin embargo, la mayoría de coeficientes estimados en la etapa de la regresión cuantílica no han sido significativos. Esto podría deberse al hecho de que la regresión cuantílica es altamente sensible a la estimación de la probabilidad θ_i^*. Aquellos coeficientes para los cuales sí se ha obtenido una significatividad estadística para la regresión pueden servirnos como información útil para identificar aquellos factores que caracterizan a las pólizas más riesgosas de la cartera de la compañía.

Con la probabilidad anteriormente nombrada, se han podido calcular los cuantiles del costo de los siniestros para cada uno de los grupos de riesgo formados. Estos

cuantiles se pueden interpretar como la pérdida a un cierto nivel de confianza θ_i^* que tendrá la compañía para cada una de las tarifas establecidas, lo cual nos da información útil sobre la manera de tarificación para aplicaciones posteriores como el cálculo de la prima pura.

Referencias

[1] Andrey A.Kudryavtsev. Using quantile regression for rate-making. *Insurance: Mathematics and Economics*, 45(2):296 – 304, 2009.

[2] Fabio Baione and Davide Biancalana. An individual risk model for premium calculation based on quantile: A comparison between generalized linear models and quantile regression. *North American Actuarial Journal*, 2019.

[3] Piet de Jong and Gillian Z. Heller. *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science. Cambridge University Press, 2008.

[4] Edward W. Frees. *Regression Modeling with Actuarial and Financial Applications*. International Series on Actuarial Science. Cambridge University Press, 2010.

[5] Antonio Heras, Ignacio Moreno, and José L. Vilar-Zanón. An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 9(2018):753–769, 2018.

[6] R. Koenker and G. W. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, January 1978.

[7] Georgios Pitselis. Quantile credibility models. *Insurance: Mathematics and Economics*, 52(3):477 – 489, 2013.

[8] Esther Portnoy. Regression-quantile graduation of australian life tables, 1946–1992. *Insurance: Mathematics and Economics*, 21(2):163 – 172, 1997.

6. Anexos

6.1. Resultados de la regresión cuantílica para cada uno de los modelos estimados.

```
Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.79048

Coefficients:
            Value     Std. Error t value  Pr(>|t|)
(Intercept)  8.16124  0.16119   50.63271  0.00000
sexeM       -0.02032  0.11716   -0.17346  0.86229
age_g1      -0.33944  0.17987   -1.88716  0.05919
age_g2      -0.15093  0.15846   -0.95249  0.34089
age_g4      -0.30463  0.44527   -0.68415  0.49391

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.77

Coefficients:
            Value     Std. Error t value  Pr(>|t|)
(Intercept)  7.96013  0.13651   58.31037  0.00000
sexeM       -0.03658  0.09666   -0.37839  0.70515
age_g1      -0.28457  0.14651   -1.94234  0.05215
age_g2      -0.11385  0.13376   -0.85116  0.39472
age_g4      -0.32183  0.40028   -0.80402  0.42142

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.71879

Coefficients:
            Value     Std. Error t value  Pr(>|t|)
(Intercept)  7.56249  0.09396   80.48287  0.00000
```

```
sexeM       -0.02302  0.07100   -0.32426  0.74575
age_g1      -0.15200  0.10087   -1.50692  0.13189
age_g2      -0.05422  0.09174   -0.59099  0.55455
age_g4      -0.16310  0.14444   -1.12921  0.25886

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.65013

Coefficients:
            Value     Std. Error t value   Pr(>|t|)
(Intercept)  7.22255   0.06799  106.22966   0.00000
sexeM       -0.00849   0.05129   -0.16548   0.86857
age_g1      -0.09199   0.07552   -1.21816   0.22321
age_g2      -0.03424   0.06885   -0.49736   0.61895
age_g4       0.02405   0.13102    0.18359   0.85434

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.75852

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  7.90644  0.12077   65.46512  0.00000
sexeM       -0.06321  0.08197   -0.77114  0.44066
age_g1      -0.26538  0.12736   -2.08362  0.03724
age_g2      -0.11352  0.12173   -0.93254  0.35110
age_g4      -0.27301  0.35498   -0.76909  0.44187

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.73393

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  7.66109  0.09938   77.08515  0.00000
sexeM       -0.02410  0.07453   -0.32339  0.74642
age_g1      -0.15819  0.11036   -1.43345  0.15178
```

```
age_g2       -0.07988  0.09921   -0.80516  0.42076
age_g4       -0.22643  0.15988   -1.41623  0.15676

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.67245

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  7.31128  0.07650   95.57161  0.00000
sexeM        0.01596  0.05692    0.28041  0.77917
age_g1      -0.11445  0.08633   -1.32567  0.18500
age_g2      -0.04539  0.07701   -0.58948  0.55556
age_g4      -0.04746  0.09762   -0.48615  0.62688

Call: rq(formula = log(cout) ~ sexe + age_g, tau = reglog$p_astec,
    data = car1)

tau: [1] 0.59004

Coefficients:
            Value     Std. Error t value   Pr(>|t|)
(Intercept)   6.94291   0.06161  112.68997   0.00000
sexeM        -0.03435   0.04680   -0.73390   0.46304
age_g1       -0.00707   0.07476   -0.09456   0.92466
age_g2        0.01950   0.06302    0.30947   0.75698
age_g4        0.12724   0.12576    1.01180   0.31167
```

6.2. Anexo - Código en SAS para la regresión logística.

```
 proc genmod data=work.base descending;
class sex_g age_g;
model claim = sex_g age_g/ dist=bin;
fwdlink link=log((_MEAN_/exposition)/(1-(_MEAN_/exposition)));
invlink ilink=exposition*exp(_XBETA_)/(1+exp(_XBETA_));
store out=REGLOG;
run;

data new;
input sex_g age_g;
cards;
1 1
1 2
1 3
1 4
0 1
0 2
0 3
0 4
;
run;
proc plm source=REGLOG;
score data=new out=preds pred=pred lclm=lower uclm=upper / ilink;
run;

data preds;
set preds;
p0=1-pred;
p_astec=(0.95-p0)/(1-p0);
run;

proc print data=preds;
var sex_g age_g p0 p_astec;
run;
```

6.3. Anexo - Código en R para la regresión cuantílica.

```
# Base de datos
car<-read.csv("baseFREQ2.csv",header = TRUE, sep = ';')
summary(car)
attach(car)

# Vector de cuantiles modificados (extraidos de la Etapa 1)
reglog<-read.csv("reglog.csv",header = TRUE, sep = ',')
head(reglog)
QM <- as.matrix(reglog)

# Regresion Cuantilica
Sexo = relevel(sexe, ref="F")
car$sexe = relevel(factor(sexe), ref="F")
car$age_g = relevel(factor(age_g), ref="4")
car1=car[car$cout>0,]
pppp=rq(log(cout)~sexe+age_g, data=car1, tau=reglog$p_astec)
summary(pppp)
```

Cuadro 11: Variables descartadas y aceptadas debido a su significación estadística.

Parámetro		Estimación	Error estándar	Pr >ChiSq
Intercept		-10.349	0.5973	0.0832
duree_permis_g	1	-0.7504	0.5806	0.1963
duree_permis_g	2	-0.8133	0.5781	0.1595
duree_permis_g	3	-0.9768	0.5781	0.0911
duree_permis_g	4	-0.6467	0.6068	0.2866
duree_permis_g	5	0.0000	0.0000	.
anne_vehicule_g	1	-178.983	7.112.985	0.9980
anne_vehicule_g	2	-182.668	16039.07	0.9991
anne_vehicule_g	3	-14.121	10.467	0.1773
anne_vehicule_g	4	-0.7950	0.1445	<.0001
anne_vehicule_g	5	-0.3649	0.0314	<.0001
anne_vehicule_g	6	0.0000	0.0000	.
marque_voiture	ALFA ROMEO	0.7469	0.1795	<.0001
marque_voiture	AUDI	0.4064	0.1617	0.0120
marque_voiture	Autres	0.3395	0.2280	0.1364
marque_voiture	BMW	0.4948	0.1612	0.0021
marque_voiture	CHRYSLER	0.2991	0.2397	0.2120
marque_voiture	DAEWOO	0.1492	0.2588	0.5644
marque_voiture	DAIHATSU	-0.3724	0.3793	0.3262
marque_voiture	FIAT	0.4149	0.1610	0.0100
marque_voiture	FORD	0.3541	0.1487	0.0172
marque_voiture	GM	0.1521	0.1482	0.3047
marque_voiture	HONDA	0.3010	0.1480	0.0421
marque_voiture	HYUNDAI	0.4357	0.2407	0.0703
marque_voiture	KIA	0.6219	0.3069	0.0427
marque_voiture	LADA	-0.4307	0.4931	0.3824
marque_voiture	LANCIA	0.5228	0.2259	0.0207
marque_voiture	MAZDA	0.3680	0.1647	0.0255
marque_voiture	MERCEDES-B	0.3322	0.1636	0.0423
marque_voiture	MITSUBISHI	0.4045	0.1695	0.0170
marque_voiture	NISSAN	0.3166	0.1540	0.0399
marque_voiture	PEUGEOT	0.2610	0.1509	0.0837
marque_voiture	PORSCHE	-169.411	6.981.325	0.9981
marque_voiture	RENAULT	0.1765	0.1478	0.2322
marque_voiture	ROVER	0.4927	0.1958	0.0119
marque_voiture	SAAB	0.3875	0.2508	0.1224
marque_voiture	SEAT	0.3704	0.1788	0.0383
marque_voiture	SKODA	0.3049	0.2288	0.1825
marque_voiture	SUBARU	-0.0471	0.2857	0.8690
marque_voiture	SUZUKI	0.0365	0.2161	0.8660
marque_voiture	TOYOTA	0.2327	0.1524	0.1268
marque_voiture	VOLKSWAGEN	0.4098	0.1450	0.0047
marque_voiture	VOLVO	0.0000	0.0000	.
sex_g	0	-0.1626	0.0318	**<.0001**
sex_g	1	0.0000	0.0000	.
age_g	1	0.4947	0.1097	**<.0001**
age_g	2	0.4126	0.0937	**<.0001**
age_g	3	0.2742	0.0884	**0.0019**
age_g	4	0.0000	0.0000	.